Advances in Knowledge-Based Technologies

Proceedings of the
Master and PhD Seminar

Winter term 2019/20, part 2

Softwarepark Hagenberg
SCCH, Room 0/2
7 February 2020

Software Competence Center Hagenberg
Softwarepark 21
A-4232 Hagenberg
Tel. +43 7236 3343 800
Fax +43 7236 3343 888
www.scch.at

Fuzzy Logic Laboratorium Linz
Softwarepark 21
A-4232 Hagenberg
Tel. +43 7236 3343 431
Fax +43 7236 3343 434
www.flll.jku.at

Program

Session 1 — Chair: Susanne Saminger-Platz

09:00 Lisa Ehrlinger:
Automating Data Quality Measurement with Tools: State-of-the-Art and Future Potential

09:30 Michal Lewandowski:
Towards a ReLU network based distance for comparing GANs with small samples

Session 2 — Chair: Bernhard Moser

10:00 Florian Sobieczky:
Some graph-representations of manufacturing process data applied to anomaly detection

Automating Data Quality Measurement with Tools: State-of-the-Art and Future Potential

Lisa Ehrlinger*[1,2]

[1]Johannes Kepler University Linz, Austria
[2]Software Competence Center Hagenberg, Austria

Abstract

High-quality data is key to interpretable and trustworthy data analytics and the basis for meaningful data-driven decisions. In practical scenarios, data quality is typically associated with data preprocessing, profiling, and cleansing for subsequent tasks like data integration or data analytics. However, from a scientific perspective, a lot of research has been published about the measurement (i.e., the detection) of data quality issues and different generally applicable data quality dimensions and metrics have been discussed. In this work[1], we close the gap between research into data quality measurement and practical implementations by investigating the functional scope of current data quality tools. With a systematic search, we identified 667 software tools dedicated to "data quality", from which we evaluated 13 tools with respect to three functionality areas: (1) data profiling, (2) data quality measurement in terms of metrics, and (3) continuous data quality monitoring. We selected the evaluated tools with regard to pre-defined exclusion criteria to ensure that they are domain-independent, provide the investigated functions, and are evaluable freely or as trial. This survey aims at a comprehensive overview on state-of-the-art data quality tools and reveals potential for their functional enhancement. Additionally, the results allow a critical discussion on concepts, which are widely accepted in research, but hardly implemented in any tool observed, for example, generally applicable data quality metrics.

This presentation will summarize the key findings from the DQ tool survey, which was held previously as invited talk[2] at the renowned MIT CDOIQ 2019 (Chief Data Officer and Information Quality Symposium) at MIT, Cambridge, MA, USA. The inventor Richard Wang is pioneer and leader in DQ research and author of several seminal publications into DQ.

*lisa.ehrlinger@jku.at
[1]`https://arxiv.org/abs/1907.08138`
[2]`https://siliconangle.com/2019/08/12/do-businesses-run-on-premium-data-new-study-assesses-variables-in-data-quality-tools-mitcdoiq-womenintech`

Towards a ReLU network based distance for comparing GANs with small samples

Michał Lewandowski

February 4, 2020

[Moser et al. 2018] proposed the way of creating the code space from activations of ReLU based network, that is assigning value one to a node with strictly positive activation value and keeping it zero otherwise. Interestingly, code space is isomorphic with a tessellated input space.

There exist numerous ways of comparing GANs, e.g. Inception Score [Salimans et al., 2016], Fréchet Inception Distance [Heusel et al., 2017], GILBO [Alemi and Fischer, 2018]. These measures are not flawless, e.g. in our experiments FID displays strong dependence on the sample size. We start with proposing a novel class of metrics on the code space, namely Wasserstein distance with one of binary metrics as a base measure, we start with Hamming distance, i.e. the number of dissimilarities between the sequences of code.

As for beginning, we are interested in (1) how much of information do we loose working with code space instead of original activation values of ReLU based network, (2) can we visually distinguish between images of different classes through dimensionality reducing embeddings such as (a) PCA, (b) t-SNE, (c) UMAP, (d)...?

With our work we aim at deepening mathematical understanding of the interdependence between a deep model represented as neural network, its induced geometry (tessellation) in the input space and the role of the decision function, further we plan to investigate possible project's extensions to Transfer Learning and leveraging transfer learning to improve distributed deep learning by means of dedicated regularization strategies.

References

[Alemi and Fischer, 2018] Alemi, A. A. and Fischer, I. (2018). Gilbo: One metric to measure them all.

[Heusel et al., 2017] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium.

[Salimans et al., 2016] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans.

Some graph-representations of manufacturing process data applied to anomaly detection

Florian Sobieczky - SCCH

Abstract:

Various graphical models are available representing time series data that focus on specific characteristics of the distribution of the underlying stochastic process [1], generalizing ideas developed as early as 1971, by Zahn [2]. We select such a representation yielding clustering via minimal spanning trees [3, 4] for the purpose of anomaly detection typically relevant in the context of waste reduction and avoidance [5]. Using some results from spectral graph theory [6], we show how to estimate the intensity of waste production using estimates of graph eigenvalues.

Literature:

[1] X. Dong, D. Thanou, M. Rabbat, P. Frossard, Learning graphs from data: A signal representation perspective, arxiv.org/abs/1806.00848.

[2] C. T.Zahn, *IEEE Trans. Comp.*, Vol C-20 (1971), No. 1, pp. 68—86.

[3] O. Grygorash, Y. Zhou, Z. Jorgensen: MST based clustering algorithms, *Proceedings of the 18th IEEE ICTAI Conference 2006*, 2006.

[4] X. Lv, Y. Ma, X. He, CciMST: A Clustering Algorithm Based on Minimum Spanning Tree and Cluster Centers, *Math. Problems in Enginieering* (2018), Article ID 8451796.

[5] M. Despeisse, F. Mbaye, P.D. Ball, A. Levers, The emergence of sustainable manufacturing practices, *Production, Planing and Control*, Vol. 23 (2012), No. 5, pp. 354—376.

[6] M. Fiedler, Absolute algebraic connectivity of trees, *Lin Multilin. Alg.*, Vol. 26 ,(2008), pp. 85—106.